Unusual Event Detection in Crowded Scenes Using Bag of LBPs in Spatio-temporal Patches

Jingxin Xu, Simon Denman, Clinton Fookes, Sridha Sridharan
Image and Video Laboratory
Queensland University of Technology
GPO Box 2434, Brisbane 4001, Australia
Emails: {j15.xu, s.denman, c.fookes, s.sridharan}@qut.edu.au

***Abstract*—Modelling events in densely crowded environments remains challenging, due to the diversity of events and the noise in the scene. We propose a novel approach for anomalous event detection in crowded scenes using dynamic textures described by the Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) descriptor. The scene is divided into spatio-temporal patches where LBP-TOP based dynamic textures are extracted. We apply hierarchical Bayesian models to detect the patches containing unusual events. Our method is an unsupervised approach, and it does not rely on object tracking or background subtraction. We show that our approach outperforms existing state of the art algorithms for anomalous event detection in UCSD dataset.**

I. Introduction

Event detection in crowded scenes is a major topic of interest in computer vision. Since object tracking is very challenging in crowded scenes, research in this field has focused mainly on extracting local motion features. Local motion features are usually based on optical flow [1, 2]. However, optical flow is often unreliable, especially for textureless regions, and visual features reflected from optical flow are very limited [3]. Recent research [4] indicates that dynamic texture is a more suitable technique. Dynamic textures are sequences of images of moving scenes that exhibit spatio-temporal stationary properties [5]. Dynamic texture is referred to as "temporal texture" in [6], which is defined as the motion patterns independent to time and space. As a result, they can be treated with statistical techniques [6]. Typical examples of dynamic texture include waves, fire, smoke, clouds, trees moving in the wind, etc. Thus this technique can be used for detection of natural disasters such as a forest fire [7], and foreground segmentation [8, 9] in some time-varying backgrounds such as waves on the water, trees in the wind and moving crowds. When applying dynamic textures for event detection in crowded environments, the scene is divided into a set of spatio-temporal patches [4, 10], where stationary properties of the motion patterns are observed.

A variety of mathematical representations of dynamic textures have been proposed. In [5] dynamic textures are modelled as auto-regressive moving average processes (ARMA). Recognition of dynamic textures [11] represented by ARMA models is generally based on discriminative methodologies. As a result, this technique requires prior labeling of normal and abnormal events for event detection. However, due to the diversity of the events that can potentially occur, it is not realistic to annotate all normal or abnormal events beforehand. Typically, these applications [1, 2, 4] require generative models to provide unsupervised learning and identify those patterns with low probabilities as abnormal. Chan and Vasconcelos [12] proposes the Mixture of Dynamic Texture (MDT) on top of the ARMA representation. In [12], a motion pattern is modelled as samples from a set of underlying dynamic textures. This model has a stronger ability to represent motion patterns compared to [5]. For instance, the motion pattern of a fire is usually co-exists with the motion pattern of smoke. More significantly, as a generative model to recognize motion patterns, it can support unsupervised learning. Mahadevan *et al.*[4] applies MDTs to detect anomalous events in crowded scenes, by considering both temporal abnormalities and spatial abnormalities. They show that their approach is more reliable than previous works [2, 13, 14] which rely on optical flow and the social force model.

Alternatively, dynamic textures can also be described by Local Binary Patterns (LBP) [15]. Traditional LBP [16] has been widely used as a 2D texture descriptor, since it is simple, efficient, robust to illumination variations and affine transformations. Zhao and Pietikainen [15] extends LBP into volume local binary patterns (VLBP), by combining the temporal information to model dynamic textures. In order to simplify the application, only the co-occurrences from Three Orthogonal Planes (TOP) are considered, thus this is called LBP-TOP. Compared to ARMA based dynamic texture, the LBP-TOP descriptor has the following benefits [15]: 1) combination of motion feature and appearance feature; 2) processing locally to catch the spatio-temporal transition information; 3) insensitivity to illuminations and affine transformations; 4) computational simplicity; 5) multi-resolution analysis.

In this paper, we propose using Latent Dirichlet Allocation (LDA) to model LBP-TOP based dynamic textures. We show our results for an unusual event detection application. We use a spatio-temporal patch architecture. The motion pattern in a patch is represented by samples from K underlying dynamic textures, where K is assumed to be known beforehand. Because LDA is a generative model, our application is able to detect anomalous events in crowded scenes through identifying low likelihood patches. We evaluate our application on the UCSD Abnormality Dataset [4], and show that our proposed

approach outperforms the Mixture of Dynamic Textures algorithm of [4] which has been shown to outperforms several other earlier algorithms.

II. Connections to related works

Current state-of-the-art algorithms [1, 2, 4, 14, 17] for unusual event detection are novelty detection applications based on extracting local motion features. In [4], it has been demonstrated that the Mixture of Dynamic textures is a better representation for unusual event detection than the optical flow, by comparing several recent algorithms [2, 13, 14]. However, the dynamic textures in [4] are modelled as ARMA and recent research in facial expression recognition [15] indicates that the LBP-TOP is a stronger descriptor for dynamic texture. Ma and Cisar [10] apply LBP-TOP based dynamic textures for event detection in a spatio-temporal framework. However, their application is for event recognition in a supervised approach, without any learning models, while our algorithm is an unsupervised approach which uses a comprehensive learning model, Latent Dirichlet Allocation. Various classifiers that can support novelty detection are able to be used. In [2], Gaussian Mixture models (GMM) are used for local anomaly detection. However, GMMs often cause an overfitting problem when the dimension of the feature vector is high, as the covariance matrix becomes singular. In [17], Hidden Markov Models (HMMs) are used in the spatio-temporal patches framework, while the observations in a hidden state are assumed to draw from the mixture of Gaussians distribution. As a result, the same overfitting problem will occur when the dimensionality of the input is high. Latent Dirichlet Allocation is a topic model based on the "bag of words" assumption. It counts the histogram of the feature elements and assumes all elements in the histogram are independent and identically distributed. As a result, LDA is able to model input histogram with very large dimensionality. Latent Dirichlet Allocation [18] models the documents as a bags of words generated by K topics, by minimizing the sum of likelihoods of all documents in the corpus. Different from other works using LDAs for activity modelling [1, 13], the proposed approach models a spatio-temporal patch rather than a short video clip as a "document", and does not use optical flow.

The remainder of the paper is organized as follows: Section III explains our algorithm in detail; Section IV presents an evaluation of our algorithm; and the paper is concluded in Section V.

III. Abnormality Detection

This section presents our proposed algorithm for anomalous event detection using LBP-TOP based dynamic textures in detail. Our algorithm contains three parts: feature extraction, model training and detection, which are explained in Section A, B and C, respectively.

A. Feature Extraction Using LBP-TOP

Figure 1 shows a local 3×3 neighborhood of a gray scale image. Texture T is defined as the joint distribution of intensities from the nine pixels [16]:

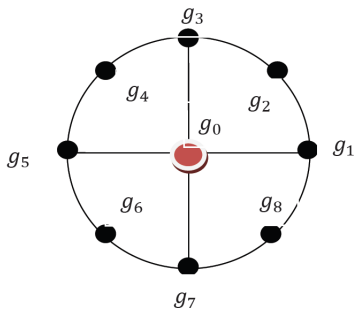

Fig. 1. Local Binary Pattern:based on the centre pixel (g_0) and its four neighbours (g_1,g_3,g_5,g_7), the intermediate pixel values (g_2,g_4,g_6,g_8) are interpolated. The joint distribution of these intensities define the texture.

$$T = p(g_0, g_1, g_2, g_3, g_4, g_5, g_6, g_7, g_8), \tag{1}$$

where $g_i (i = 0, ..., 8)$ are the intensities of the pixels, and g_2,g_4,g_6,g_8 are computed by interpolation. We can subtract g_0 from the eight surrounding pixels' intensities without losing information [16]:

$$\begin{aligned} T = p(&g_0, g_1 - g_0, g_2 - g_0, g_3 - g_0, g_4 - g_0, \\ &g_5 - g_0, g_6 - g_0, g_7 - g_0, g_8 - g_0). \end{aligned} \tag{2}$$

If g_0 is assumed to be independent to the difference $g_i - g_0$, $p(g_0)$ can be extracted from Eq.(2). Since $p(g_0)$ is unrelated to local image texture, we can ignore it. Then the texture T is solely determined by the joint distribution of differences $g_i - g_0$, where $i = 1, \cdots, 8$. Since the sign of $g_i - g_0$ is invariant to gray scale changes, [16] defines the gray scale invariance local binary pattern (LBP) by considering only the sign of differences, as

$$LBP_8 = \sum_{i=1}^{8} s(g_i - g_0) 2^{i-1}, \tag{3}$$

where

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}. \tag{4}$$

This representation of LBP can be further extended to support rotation invariance. However, as our target application uses a single stationary camera, we don't require this extension. Over a local region (typically much larger than 3×3), the histogram of LBPs can be used to represent the texture.

Dynamic texture extends the traditional spatial texture into the temporal domain. Correspondingly, [15] extends the LBP into a spatio-temporal volume to model dynamic textures. Let $P(x_c, y_c, t_c)$ be the centre pixel in a spatio-temporal neighbourhood. The volume LBP (VLBP) is defined as the joint distribution of the intensities of $3 \times P + 3$ pixels on the current frame, t_c, the previous frame, $t_c - L$, and the next frame, $t_c + L$ in

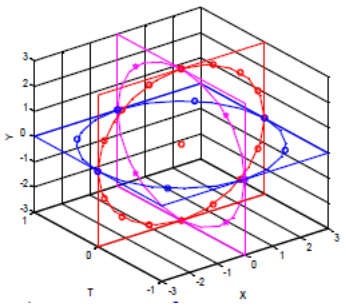

Fig. 2. LBP-TOP [15]. The three orthogonal planes are XY (red), XT (blue), and YT (purple).

$$VLBP(x_c, y_c, t_c) = \sum_{q=0}^{3P+1} s(g_q - g_c) \times 2^q, \qquad (5)$$

where P is the number of neighbours in each frame, and L is the temporal interval, g_q is neighbour pixels' intensities, and g_c is the centre pixel intensity.

In order to reduce the total number of patterns, [15] further simplifies this model, only calculating the local binary patterns from three orthogonal planes (LBP-TOP) (See Figure 2). LBPs are computed with the histogram of the output in each plane. Then the three histograms are concatenated into a single histogram.

For the application of anomalous event detection, we partition the scene into spatio-temporal patches. Within each patch, LBP-TOP is extracted. In each plane we use the 8 pixel neighbourhood. As a result, each plane contains 2^8 local binary patterns. Among the three planes, XY contains rich appearance features. XT and YT contains the motion features with limited appearance features. Similar to [19], only the XT and YT are considered in our application to make it robust to human appearance. The size of the histogram in our application is 512 bins. It should be noted that, due to the learning model used in our application (see Section B), we use the non-normalized histogram, which is different from [15].

B. Training Process

In this representation (LBP-TOP), the space-time relationships within a patch is ignored. Dynamic textures are modelled as "bag of LBPs". A generative model for the "bag of features" problem is Latent Dirichlet Allocation [18], which is used as the learning model in our application.

Latent Dirichlet Allocation (LDA) is a hierarchical Bayesian model originally proposed in natural language processing. In this model, a corpus is considered as a collection of documents, where each document is a combination of topics selected from the total K topics, where the topic draws from Dirichlet distribution. Each topic is a multinomial distribution of words in a vocabulary. LDA learns the topics in each document in an unsupervised way, and can be used to learn the likelihood of a document as well.

In our application, the entire video is a "corpus", and a spatio-temporal patch is a "document". We assume the motion

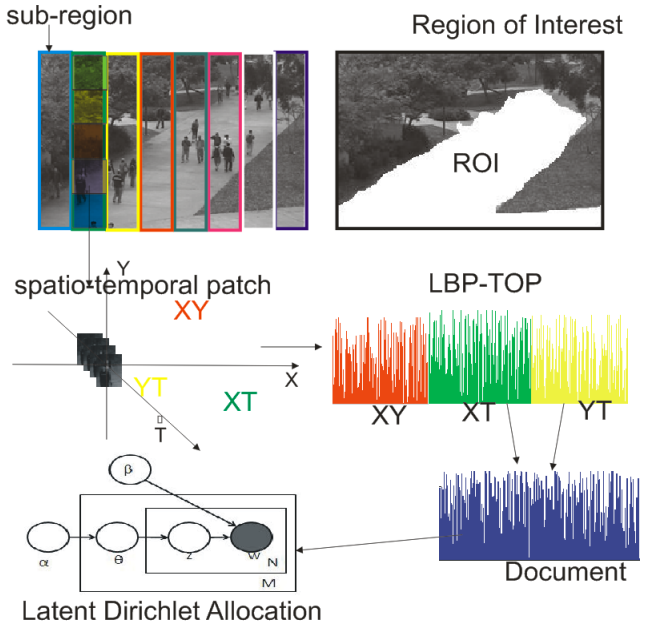

Fig. 3. Learning Process: The scene is partitioned into several sub-regions, where each sub-region has its own LDA classifier. Spatio-temporal patches are extracted from each sub-region. If a patch falls into the region of interest (ROI) we created, the LBP-TOP features are extracted.

pattern in a spatio-temporal patch is a combination of dynamic textures from the set of K topics. The vocabulary is the local binary patterns in XT and YT planes, with a fixed size of 512. This structure is similar to the Mixture of Dynamic Textures [12], where the dynamic textures are represented by ARMA model. However, since the dynamic textures are represented by LBP-TOP, all benefits from this representation such as illumination invariance and computational simplicity (see Section I) can be obtained in our approach.

Figure 3 shows the training process. We create a region of interest in the scene where features are extracted. This ROI includes the footpath where people are observed to walk, we disregard the gardens to improve computational efficiency as there are no events within these regions. The scene is further partitioned into sub-regions , where each sub-region has its own LDA classifier. If we use a single LDA classifier for the whole scene, it can work as well. However, this results in a much larger number of topics, which reduces the computational efficiency both in the training and detection processes. In each sub-region, spatio-temporal patches are extracted, and the LBP-TOP features as described in Section A are extracted. The LDA models are trained using the outputs of the LBP-TOP descriptor (the non-normalized histogram of volume local binary patterns from XT and YT planes), with the number of topics manually set. The patches in the training process are non-overlapping in the spatial domain, but overlapping in the temporal domain. One could use overlapping patches in both spatial and temporal domain, or even use a sliding window approach in time. The strategy adopted in our application is simply to reduce the time taken to train the models.

C. Detection Process

Once the parameters of the LDA models are learned, they can be used to compute the likelihood of new observations. During testing, we perform the same scene division as used in the training process (see Section B). However, in the detection process we apply temporal sliding windows of spatio-temporal patches to allow us to test all the frames. The current frame is the centre frame in a patch. For each patch, a LBP-TOP histogram is generated as in Section A. The LDA model that corresponds to the location of the patch is used to calculate the log-likelihood of the observed sample. If this is lower than a threshold, an alarm is fired at the location of the patch in the centre frame. This means that the alarm will always be delayed by $\frac{\tau}{2}$, where τ is the temporal size of the patch.

IV. Experiments

We use the UCSD Abnormality Dataset [4] for evaluation [1]. The UCSD datasets contains videos of two pedestrian scenes from a campus, which are Pedestrian 1 dataset and Pedestrian 2 dataset. There are 34 training sequences and 36 test sequences in Pedestrian 1 dataset, and there are 16 training sequences and 12 test sequences in Pedestrian 2 dataset. The training datasets only contain normal events. Examples of anomalies includes a bus, a wheelchair, a bicycle, and a skater. We show several test results in Figure 4, with the anomalous events highlighted in red. Figure 5 illustrates examples of false alarms from Pedestrian 1[2] and Pedestrian 2 datasets. The real anomalous events in those images are the skater and wheelchair enclosed in blue.

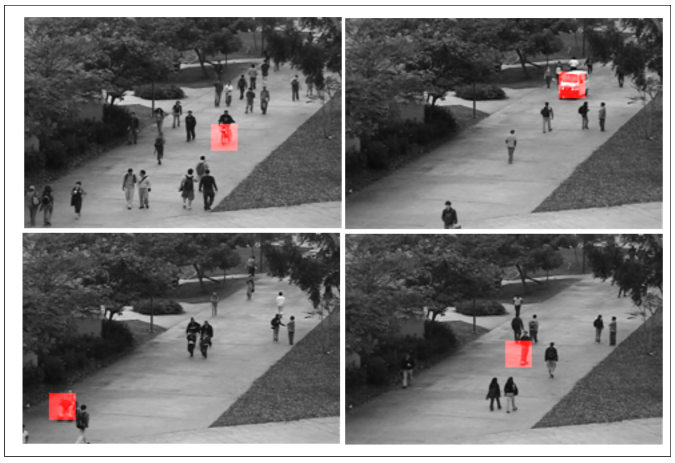

Fig. 4. Example of Anomalous Events: bicycle, bus, wheelchair, skater

The UCSD dataset contains frame level groundtruth and pixel level groundtruth. Correspondingly, we evaluate our algorithm on both frame level and pixel level. The frame level groundtruth identifies the frames containing anomalous events, while ignoring the locations of those events. The pixel level groundtruth clearly identifies the locations of the anomalous

[1]available at http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm

[2]There are several corrupted frames in 4 test cases and those test cases have been removed in our experiments.

Fig. 5. Examples of False Alarms: The skater and wheelchair enclosed in blue are the real anomalous events.Left: Pedestrian 2; Right: Pedestrian 1.

events in each frame. For example, in Figure 5, the alarms are fired at wrong locations (shown in red). However, since the frames shown in the two examples contain anomalous events, these false alarms will be calculated as true positive alarms at the frame level. For this reason, we argue that the pixel level evaluation is more appropriate, since it reflects the detection at the correct location. In order to compare algorithms, we use the evaluation method presented in [4]: it is still a frame based decision, but with a constraint that a frame is recognized as abnormal only if at least 40% of the abnormal pixels are detected.

Table I and II show the results using frame level groundtruth and pixel level groundtruth respectively, for our proposed algorithm, and for [4]. Figure 6 shows the ROC curves. We include the ROC curve of [4] with the author's permission for comparasion [3]. For the location based detection rate using the pixel level groundtruth, our algorithm achieves a 55.15% detection rate at the EER in the Pedestrian 1 dataset, which outperforms results in [4] achieving detection rate of 45% (see Table II). Since [4] does not provide the location level detection rate for Pedestrian 2 dataset, it is not able to be compared.

Our algorithm works better to detect a bicycle and a bus, than a wheelchair or a skater. This is because the motion patterns and appearance from a skater, for instance, are similar to those of walking persons (See Figure 5). False alarms are often caused when the window contains motions from two different people. The example illustrated in the left image of Figure 5 is of this kind. From the frame based ROC curve, it is seen that our approach has a better performance in Pedestrian 2 than Pedestrian 1 dataset. The reason is in the Pedestrian 1 dataset, there is significant perspective distortion. This results in individuals far smaller than the window size at the top of the image, often resulting in missed alarms. In the Pedestrian 2 dataset, there is very little perspective distortion and the performance of the algorithm improves. These problems relate to the "bags of words" assumption and can be overcome through multi-resolution analysis, which has been widely adopted in scene categorization [20].

It should be noted that, the evaluation criterion used in the pixel level groundtruth proposed in [4] has its limitations. In this criterion, the frame will be identified as abnormal if at

[3]Presentation slides of [4] is available at http://videolectures.net/cvpr2010_mahadevan_adcs/.

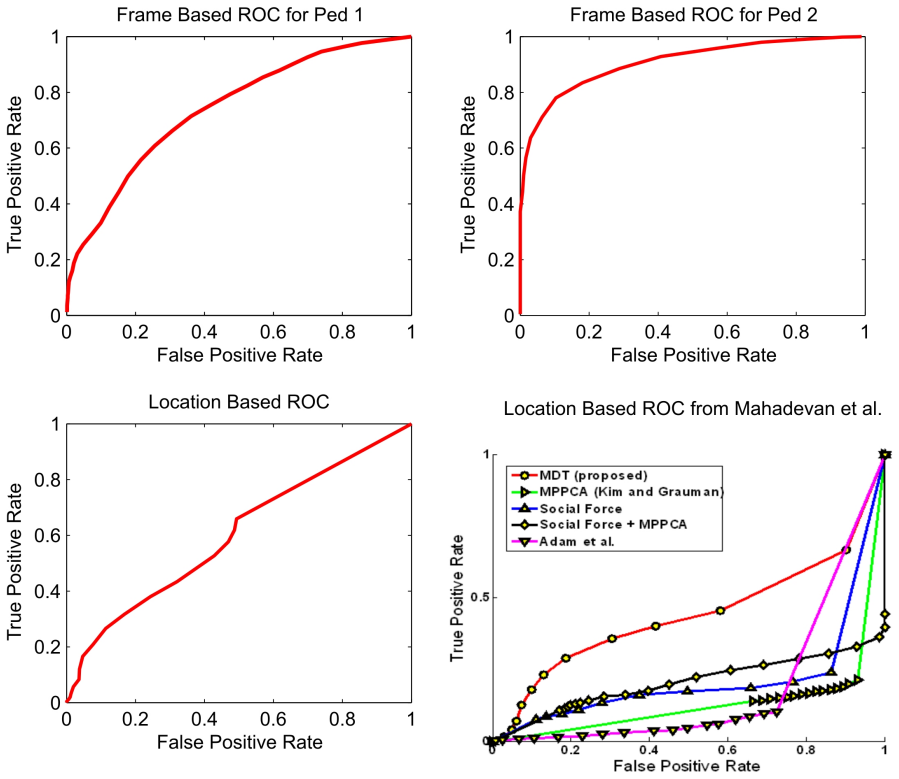

Fig. 6. ROC curves. Top: Frame based ROC curves for our approach; Down Left: Location ROC curve for our approach; Down Right: Location Based ROC from [4].

	SF[4]	MPPCA [4]	SF-MPPCA [4]	Adam et al.[4]	MDT[4]	Our approach
Ped1	31%	40%	32%	38%	25%	32.25%
Ped2	42%	30%	36%	42%	25%	17.2%
Average	37%	35%	34%	40%	25%	24.7%

TABLE I
THE EERS USING FRAME BASED GROUNDTRUTH FOR PED 1 AND PED 2

	SF[4]	MPPCA [4]	SF-MPPCA [4]	Adam et al.[4]	MDT[4]	Our approach
Ped1	21%	18%	28%	24%	45%	55.15%

TABLE II
LOCATION BASED DETECTION RATE AT EER FOR PEDESTRIAN 1 DATASET

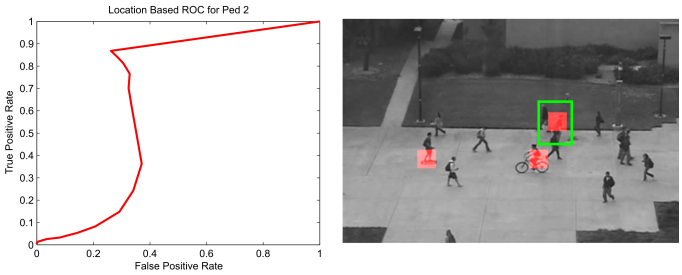

Fig. 7. Limitations of the Pixel Level Evaluation Method proposed in [4]. Left: ROC curve for Pedestrian 2. Right: an example of two true alarms (the skater and the bicycle) and a false alarm (enclosed in green) occurring in the same frame.

least 40% of abnormal pixels are sucessfully detected. When there are multiple alarms in one frame (see the right image in Figure 7), this is a true positive detection if only one of them is located at the correct location. This results in an inaccurate count of the alarms, and can lead to a particularly misleading false alarms count, as any false alarm that is present in the same frame as a true alarm is missed. This problem is more pronounced in the Pedestrian 2 dataset as the abnormal events have a longer duration than in Pedestrian 1. Our future work will investigate a more appropriate evaluation criterion.

The patches are $20 \times 20 \times 11$ size. The interval of two successive patches in time is 9 frame. Our proposed algorithm processes each frames for detection in an average time of 1.385 seconds using an Intel dual core CPU (1.96GHz and 3.33GHz) and 3.46GB RAM PC, and C++ implementation [4] [5]. In comparison, [4] requires 25 seconds to process each frames using a 3GHz CPU and 2GB RAM PC, however it is unclear what platform their system is implemented in.

[4]Source code of LDA [18] is available at http://www.cs.princeton.edu/˜blei/lda-c/ index.html

[5]Source code of LBP-TOP [15] is available at http://www.ee.oulu.fi/˜gyzhao/

V. Conclusions and Future Works

In this paper, we have proposed a novel approach, applying Latent Dirichlet Allocation to model LBP-TOP based dynamic textures to detect abnormal events. Our proposed approach retains the functionality of MDT while preserving the benefits of the LBP-TOP descriptor. The algorithm proposed in [4] has shown to outperform several state of the art algorithms. In this paper, we show that our proposed approach outperforms [4] for anomalous event detection in crowded scenes. Although we only discuss event detection in this paper, this combination of LBP-TOP and LDA can be applied to various applications where MDTs have also been applied, including video clustering and motion segmentation, while preserving the benefits of LBP-TOP.

Compared to the ARMA based dynamic texture, the LBP-TOP based dynamic texture has a lot of benefits for future extensions. The LBP-TOP features can be easily combined with other "bags of words" features, such as 3D-SIFT, histogram of oriented gradients or even colour histograms. By concatenating such features, we can learn them using a single model as outlined in this paper. It also enables the concatenation of LBP-TOP histograms from different location regions into a single histogram to model global interactions. The primary limitation of this approach is the "bag of words" assumption. That is, the spatio-temporal order of the LBP-TOP patterns in a spatio-temporal patch has been ignored. However, this problem could be overcome through multi-resolution analysis. These possibilities will be investigated in our future work.

References

[1] X. Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," vol. 31, no. 3, pp. 539–555, 2009.

[2] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 3, pp. 555 –560, march 2008.

[3] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer, 2011.

[4] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1975–1981.

[5] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," *International Journal of Computer Vision*, vol. 51, pp. 91–109, 2003, 10.1023/A:1021669406132. [Online]. Available: http://dx.doi.org/10.1023/A:1021669406132

[6] R. Polana and R. Nelson, "Detecting activities," in *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR '93., 1993 IEEE Computer Society Conference on*, Jun. 1993, pp. 2 –7.

[7] B. U. Toreyin, Y. Dedeoglu, A. E. Cetin, S. Fazekas, D. Chetverikov, T. Amiaz, and N. Kiryati, "Dynamic texture detection, segmentation and analysis," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, ser. CIVR '07. New York, NY, USA: ACM, 2007, pp. 131–134. [Online]. Available: http://doi.acm.org/10.1145/1282280.1282304

[8] J. Zhong and S. Sclaroff, "Segmenting foreground objects from a dynamic textured background via a robust kalman filter," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003, pp. 44 –50 vol.1.

[9] V. Mahadevan and N. Vasconcelos, "Background subtraction in highly dynamic scenes," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1 –6.

[10] Y. Ma and P. Cisar, "Event detection using local binary pattern based dynamic textures," in *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, 2009, pp. 38 –44.

[11] P. Saisan, G. Doretto, Y. N. Wu, and S. Soatto, "Dynamic texture recognition," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 2, 2001, pp. II–58 – II–63 vol.2.

[12] A. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 5, pp. 909 –926, May 2008.

[13] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 935 –942.

[14] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 2921 –2928.

[15] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 915 –928, 2007.

[16] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971 –987, Jul. 2002.

[17] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, june 2009, pp. 1446 –1453.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, March 2003. [Online]. Available: http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993

[19] V. Kellokumpu, G. Zhao, and M. Pietikinen, "Human activity recognition using a dynamic texture based method," in *In BMVC*, 2008.

[20] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, 2006, pp. 2169 – 2178.